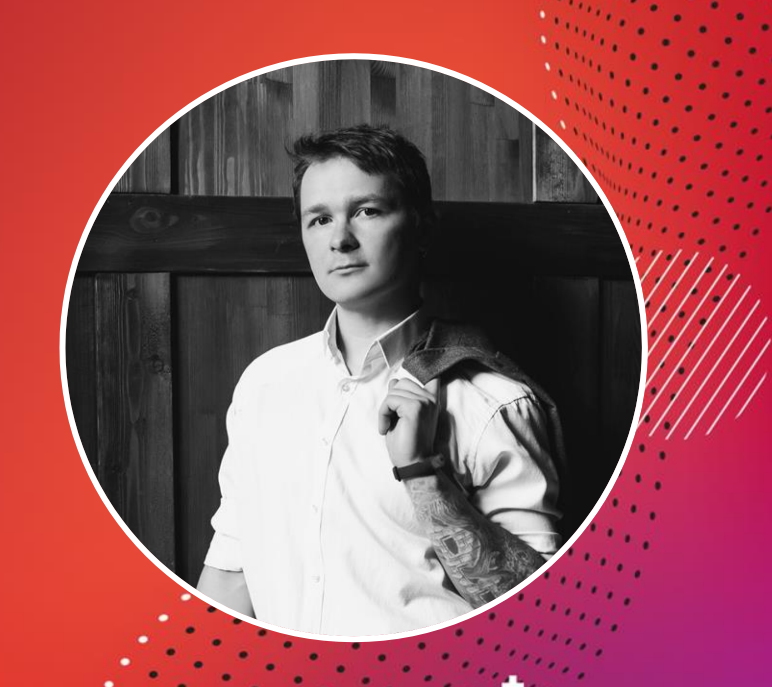
Data-centric MLOps

DMITRII EVSTIUKHIN
SENIOR SOLUTIONS ARCHITECT
@PROVECTUS









О чем поговорим

- MLOps 101: что, зачем и как
- Почему данные важнее, чем мы думали
- Инструменты и примеры решений







Провектус — профессионалы в сфере машинного обучения и искусственного интеллекта



Основана в 2010
Главный офис в Palo Alto



520 сотрудников и растем



Офисы в США, Канаде, Европе и Латинской Америке



Работаем как со стартапами, так и с корпорациями





Наши клиенты





































Вопрос к аудитории

Кто вы в компании?

1. Data Scientist / Аналитик
2. ML-Инженер
3. Data-Инженер
4. QA-Специалист
5. DevOps-Инженер
6. Менеджер





MLOps





MLOps — практически ДевОпс, но сложнее

DevOps	MLOps
Методология разработки ПО и взаимодействия инженеров различных специальностей	Методология разработки ПО и взаимодействия инженеров различных специальностей
Уже выделена отдельная профессия — DevOps-инженер	Есть ли специальная роль для того, кто этим занимается?
Относительно простой и прямолинейный процесс	Более сложный многокомпонентный процесс с бóльшим количеством участников





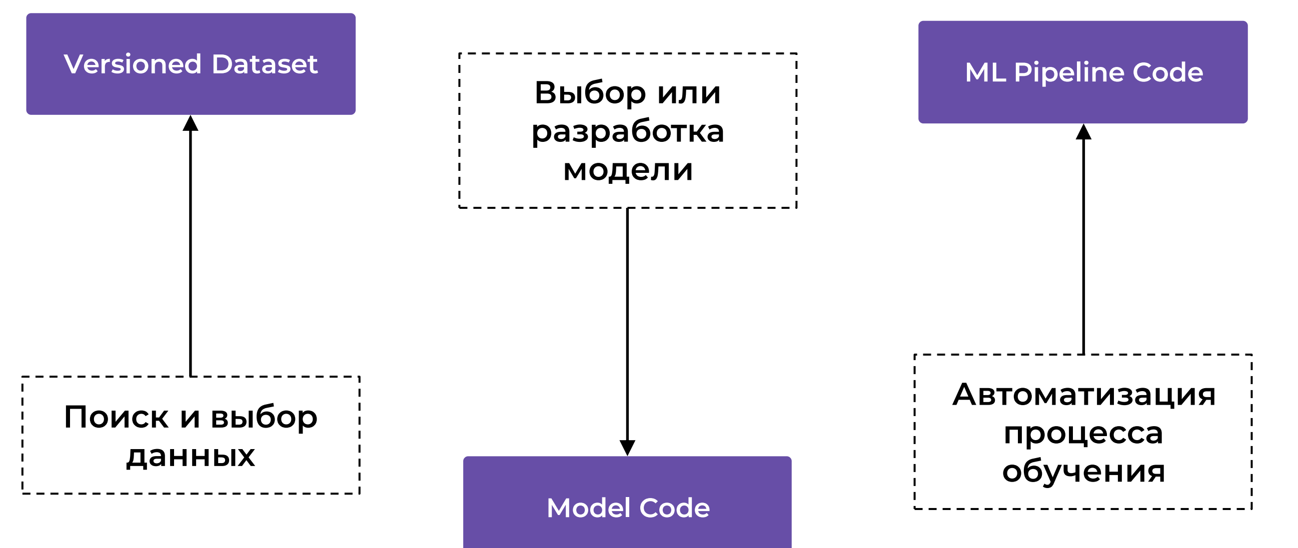
Входы MLOps

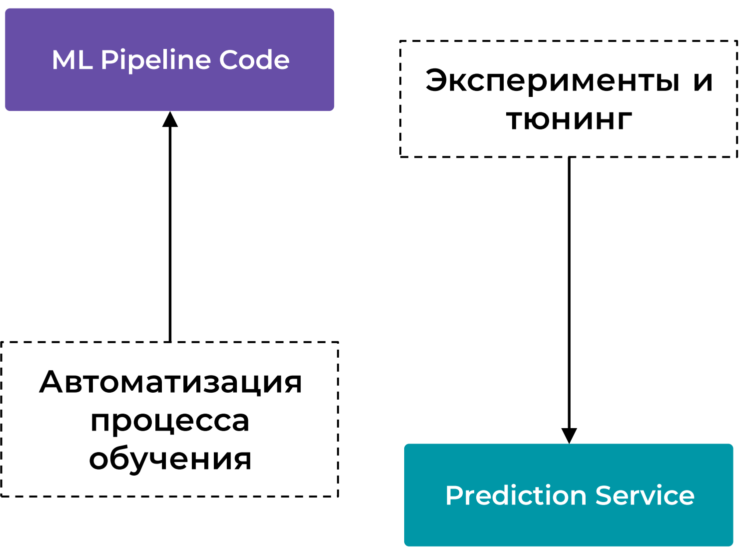
Model Code	Код модели, препроцессинга, инференса
ML Pipeline Code	Код пайплайна для оркестрации процесса обучения
Infrastructure as Code	Код инфраструктуры, конфигурация платформы
Versioned Dataset	Данные





Итерация MLOps









MLOps-пайплайн

git → Model Code
git → ML Pipeline Code
Infrastructure as Code
Feature Store → Versioned Dataset

Orchestration: Idempotent Execution

ML Metrics
Alerts
Reports
Automated Pipeline Execution
Pipeline Metadata

Production Metrics & Alerts
Model Artifacts
Prediction Service

Feedback Loop for Production Data

Профессия MLOps

ML engineer/ Data Scientist	DevOps engineer	Software Engineer	Domain Expert
55.70%	13.30%	6.00%	25.00%

60.00%
40.00%
20.00%
0.00%



*from: A Chat with Andrew on MLOps: From Model-centric to Data-centric AI
https://www.youtube.com/watch?v=06-AZXmwHjo



Вопрос к аудитории

Самое главное в решении ML-проблемы — это выбрать удобный фреймворк и построить правильную модель. Так?





Данные





Данные

- Качество данных
- Прозрачность данных
- Инструменты и примеры решений







Каскады данных

Problem Statement
Data collection and labeling
Data analysis and cleaning
Model selection
Model training
Model evaluation
Model deployment

Interacting with physical world brittleness
Inadequate application-domain expertise
Conflicting reward systems
Poor cross-organizational documentation

Impacts of cascades
Abandon / re-start process

** from "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI",*

N. Sambasivan et al., SIGCHI, ACM (2021)





IEEE Bulletin, март 2021

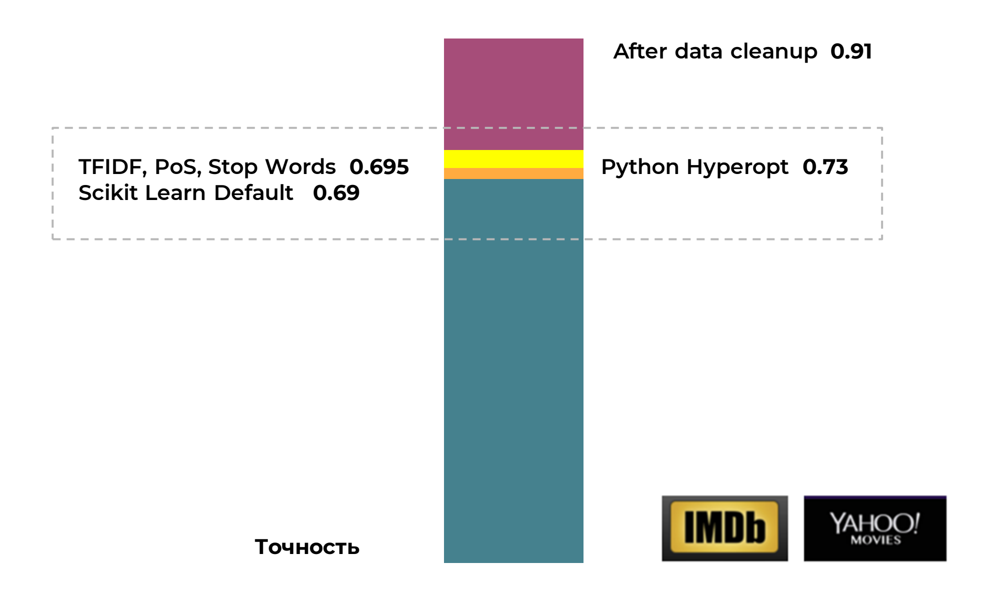
Special Issue on Data Validation for Machine Learning Models and Applications

**from: "The Bulletin of the Technical Committee on Data Engineering"*
http://sites.computer.org/debull/A21mar/issue1.htm





Эффект качественной очистки данных



Sigmod2016
Sanjay Krishnan (UC Berkeley)
И Jiannan Wang (Simon Fraser U.)
https://sigmod2016.org/sigmod_tutorial1.shtml





From Model-centric to Data-centric AI

	Обнаружение дефектов стали	Солнечные панели	Инспекция поверхностей
Базовый показатель	76.2%	75.68%	85.05%
Model-centric	+0% (76.2%)	+0.04% (75.72%)	+0% (85.05%)
Data-centric	+16.9% (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

**from: "A Chat with Andrew on MLOps: From Model-centric to Data-centric AI"
https://www.youtube.com/watch?v=06-AZXmwHjo*





GoCheck Kids Case Study

	До	После Data QA
precision	32%	40%
recall	89%	91%
false alarm rate	19%	17%
PR AUC	57%	76%



Плохие данные → Шикарная модель🤩 → Плохие результаты😫





Вопрос к аудитории

Как вы находите нужный датасет в своей организации?

1. Спрашиваете коллег
2. Используете Wiki/документацию
3. Реверс-инженерите пайплайны данных
4. Скан регэкспами по всем файлам
5. Используете Каталог данных





Управление данными:

1. Данные разбросаны по разным системам хранения: RDMS, DWH, Data Lakes, Blobs
2. Не всегда очевидно, кто владеет данными
3. Требования к данным и SLA этих данных — не определены
4. 90% команд, работающих с данными, жалуются на проблемы с поиском и доступом к данным
5. Такие команды тратят 25-50% времени только на поиск и оценку найденных данных
6. Если и есть способ показать всю историю происхождения данных, то такая история не касается области ML







Рынок каталогов данных

Product	OSS	Self-hosted	Search based	Lineage based	Network based	Federation	UX Personalization	AI Autowiring features	ML First citizen	Data QA Int	Profiling
Open Data Catalog	**OSS**	Yes	Yes	Yes	Yes	Yes	**Roadmap**	**Roadmap**	Yes	**Roadmap**	**Roadmap**
Amundsen (Lyft)	**OSS**	Yes	Yes	Yes	Yes	**No**	**No**	**No**	**No**	**Roadmap**	**No**
Datahub (LinkedIn)	**OSS**	Yes	Yes	Yes	Yes	**No**	**No**	**No**	**No**	**Roadmap**	**Roadmap**
Marquez (WeWork)	**OSS**	Yes	Yes	Yes	**No**	**Roadmap**	**No**	**No**	**No**	**No**	**No**
Magda	**OSS**	Yes	Yes	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Apache Atlas (Hortonworks)	**OSS**	Yes	Yes	Yes	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Collibra Data Catalog	**Prop**	Yes	Yes	Yes	N/A	**No**	Yes	N/A	**No**	N/A	N/A
Alation Data Catalog	**Prop**	Yes	Yes	Yes	**No**	**No**	Yes	**No**	**No**	Yes	**No**
Atlan	**Prop**	Yes	Yes	Yes	**No**	**No**	N/A	N/A	**No**	N/A	N/A
Informatica Data Catalog	**Prop**	Yes	Yes	Yes	Yes	**No**	N/A	N/A	**No**	Yes	Yes
Data World	**Prop**	**No**	Yes	Yes	N/A	**No**	N/A	N/A	**No**	N/A	N/A
Talend	**Prop**	Yes	Yes	Yes	N/A	**No**	Yes	N/A	**No**	Yes	Yes
Datakin	**Prop**	Yes	Yes	Yes	**No**	**No**	**No**	**No**	**No**	**No**	**No**
Zeenea Data Catalog	**Prop**	**No**	Yes	Yes	N/A	**No**	N/A	N/A	**No**	N/A	N/A
Google Cloud Data Catalog	**Cloud**	**No**	Yes	Yes	**No**	**No**	N/A	N/A	**No**	N/A	N/A
Azure Data Catalog	**Cloud**	**No**	Yes	Yes	N/A	**No**	N/A	N/A	**No**	N/A	N/A
Monte-Carlo	**Prop**	**No**	Yes	Yes	**No**	**No**	**No**	**No**	**No**	**No**	Yes
Metaplane	**Prop**	N/A	Yes	Yes	N/A	**No**	N/A	N/A	N/A	N/A	N/A
Ataccama	**Prop**	Yes	Yes	Yes	**No**	**No**	Yes	**No**	**No**	Yes	Yes

Что делать?





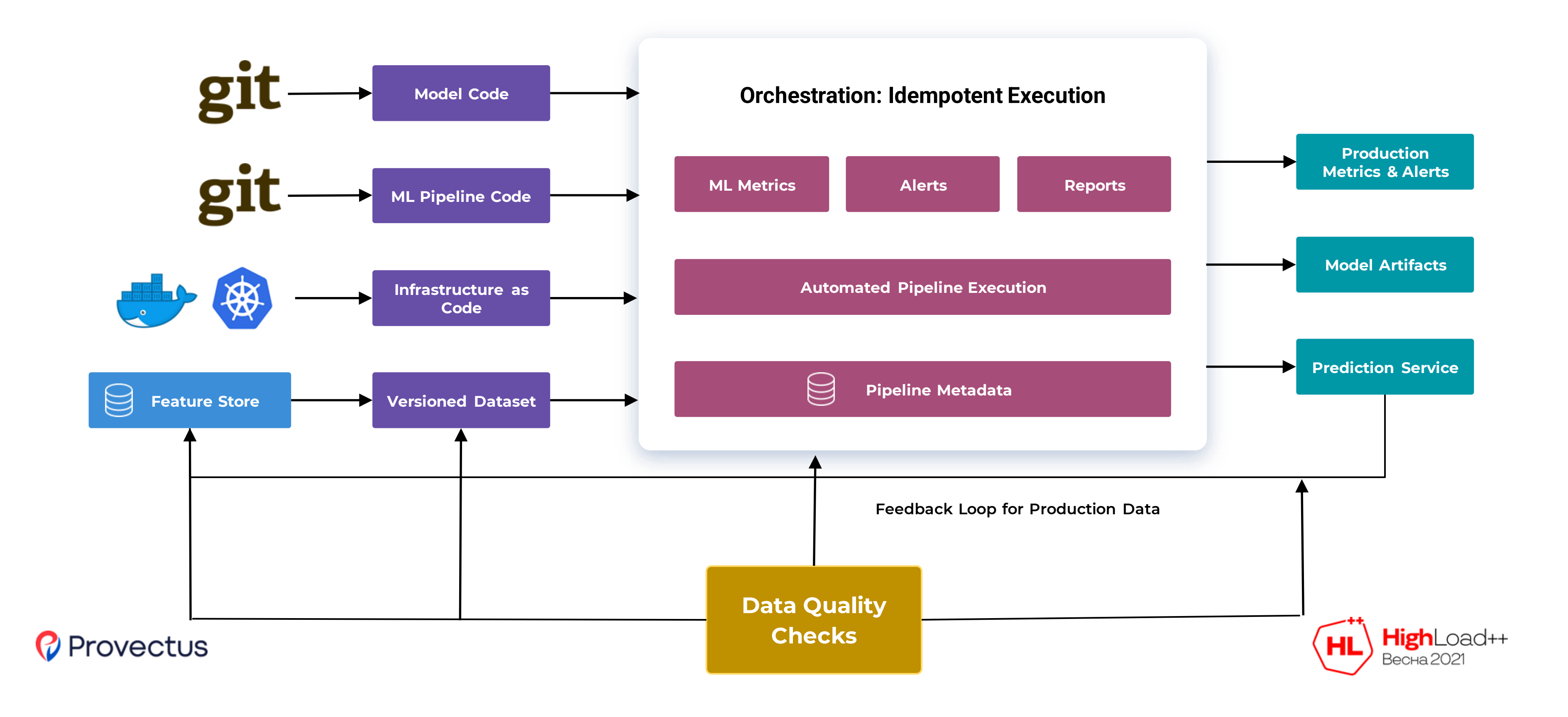
Что же делать и как?

- Тестировать данные
- Завести себе хороший Каталог данных
- Думать о данных как о самостоятельном продукте





Качество данных в MLOps-пайплайне





Versioned Dataset

ML Model

Deployed Model



Outlier Detector

Dataset Rules Validation

Model Validation

Regression Test

Kubeflow

Dataset Bias Checker

Model Test for Bias

Business Acceptance

Statistical Assertions

Model Security Test

Traffic Replay





Что тестировать в данных

Стандартные проверки данных:

- Дублирование
- Пропущенные значения
- Синтаксические ошибки
- Ошибки форматирования
- Семантические ошибки
- Целостность

Продвинутые методы:

- Проверки распределения
- Критерий Колмогорова-Смирнова
- Критерий хи-квадрат
- Автоматический поиск аномалий
- Автоматическая генерация ограничений





Менее абстрактный пример Data QA

Streaming Data Quality Analyzer

kafka

AWS Kinesis

Azure Event Hub

Spark Streaming

Streaming Data

DELTA LAKE

Bronze Layer

Down stream systems

Data Quality Checks

DELTA LAKE

Data Quality Metrics

966 13.19M 369.50K

Daily/ Hourly Quality Metrics





from: https://databricks.com/blog/2020/03/04/how-to-monitor-data-stream-quality-using-spark-streaming-and-delta-lake.html

Чем тестировать

- Deequ
  https://github.com/awslabs/deequ
- Great Expectations
  https://greatexpectations.io
- Tensorflow Data Validation
  https://www.tensorflow.org/tfx/data_validation/get_started
- DBT
  https://blog.getdbt.com/data-testing-framework/





Где искать

- Data Mesh — отдельные каталоги данных отдельных продуктов
- Федерация каталогов и централизованный каталог
- Все(!) метаданные в одном месте





Спецификация Open Data Discovery

Before ODD Spec

Data Sources

Feature Stores	ETL Tools	ML Pipelines	Warehouses	Data Quality

Amundsen	DataHub	Any other

Data Catalogs

With ODD Spec

Data Sources

Feature Stores	ETL Tools	ML Pipelines	Warehouses	Data Quality

ODD Spec

Amundsen	DataHub	Any other

Data Catalogs





Any Data Catalog

ODD Adapter API	ODD Ingestion API	Puller

Any Data Catalog

ODD Adapter API	ODD Ingestion API	Puller

ODD Adapter API

- Kubeflow Adapter
- AWS Glue Adapter

Any data sources

- Airflow
- Kubeflow
- AWS Glue





Каталог данных Open Data Discovery

1. Реализует Open Data Discovery-спецификацию
2. Глобальный федеративный каталог данных для поиска
3. Строится с поддержкой ML-сущностей, Data Quality, Data Lineage
4. Составная архитектура для удовлетворения требований вашей стратегии работы с данными и определенных бизнес-требований
5. Open-source для более простой интеграции с существующими инструментами





ML-Сущности

OpenDataDiscovery

Employee

Lena Mikheeva

Filters

Datasource 3

Data Consumers type 2

- [x] ML Model (28K)
- [ ] Dashboard (11K)

Namespace 19

Owner 78

Tags 1K

Datasets 29K | Transformers 12K | **Data Consumers 215** | Quality Tests 236 | Data Inputs 1K

ML_model_superstore_master — 22 days ago

Sources: superstore_dev

Namespace	Datasource	Created
Finance Department	DHS	27 Mar 2019 by Elizabeth Smith

sales · employee · staff · finance employees · 2020

ML_model_sales_overview — 1 months ago

Sources: sales_data-base_2019

Namespace	Datasource	Created
Sales	Civic Education Study Survey	27 Mar 2019 by Robert Vanstain

sales · employee

ML_model_employee_reports — 1 year 7 months ago



Качество данных

OpenDataDiscovery

Search

Lena Mikheeva

Quality Tests > test_superstore_master

test_superstore_master

27 days ago

Overview | Structure | Lineage

Dataset | 3 superstore_dev, sales_short-term_projects, finance_yearly_reports

Suit Url | 1 master_dev_jebus_cross_project

Namespace	Datasource	Created	Owner
Finance	Civic Education Study Survey	7 Mar 2019	Jeff Millborn

Metadata

CUSTOM

Database name	avroshmavro
Storage_descriptor.input_format	org.apache.hadoop.hive.ql.io.avro.AvroContainerInputFormat
☝ Storage_descriptor.output_format	org.apache.hadoop.hive.ql.io.avro.AvroContainerOutputFormat
Storage_descriptor.compressed	false
☝ Storage_descriptor.number_of_buckets	0
Storage_descriptor.serde_info.serializati...	org.apache.hadoop.hive.serde2.avro.AvroSerDe

View all (21)

PRE-DEFINED

Database name	avroshmavro

Tags

sales, employee, staff, finance employees, marketing, pepsico brand, yearly reports, 2020, policy agreement, leagal department, updated

Test Report

Score 89%

Tests 41

Passed	34
Failed	4
Broken	2
Skipped	1
Unknown	0

Show All



Data-centric MLOps

Data Observability Platform

git → Model Code →

git → ML Pipeline Code →

Infrastructure as Code →

Feature Store → Versioned Dataset →

Orchestration: Idempotent Execution

ML Metrics | Alerts | Reports

Automated Pipeline Execution

Pipeline Metadata

→ Production Metrics & Alerts

→ Model Artifacts

→ Prediction Service

Feedback Loop for Production Data

Data Quality Checks





Заключение

- MLOps — это не роль и не профессия, это процесс
- Качество данных решает
- Способность найти качественные данные решает не меньше
- Provectus участвует в решении последней проблемы







Спасибо!

Q&A

Dmitrii Evstiukhin
LinkedIn: linkedin.com/in/devstiukhin
Telegram: @Myafk
Email: devstiukhin@provectus.com



